

Claims

1. A file server system including  
a first file server including a file server change memory;  
a second file server including a file server change memory;  
a mass storage element;  
said first file server and said second file server being coupled to said mass  
storage element;

means for copying a descriptor of a file system change to both said first and  
second file servers, whereby said first file server processes said file system change while  
said second file server maintains its copy of said descriptor in its file server change mem-  
ory; and

means for said second file server to perform a file system change in its file  
server change memory in response to a service interruption by said first file server.

2. A system as in claim 1, including at least one said mass storage ele-  
ment for each said file server.

3. A system as in claim 1, wherein a first said file server is disposed for  
processing said file system changes atomically, whereby a second said file server can on  
failover process exactly those file system changes not already processed by said first file  
server.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21

4. A system as in claim 1, wherein a first said file server is disposed to respond identically to service interruptions for itself and for a second said file server.

5. A system as in claim 1, wherein at least one said file server is disposed to delay output to said mass storage element without delaying a response to file system changes.

6. A system as in claim 1, wherein at least one said file server responds to a file system change before committing a result of said file system change to mass storage.

7. A system as in claim 1, wherein  
each one of said file servers is coupled to at least a portion of said file server change memory using local memory access; and  
each one of said file servers is coupled to at least a portion of said file server change memory using remote memory access

8. A system as in claim 1, wherein said descriptor includes a file server request.

1           9.     A system as in claim 1, wherein said file server change memory in-  
2 cludes a disk block.

3  
4           10.    A system as in claim 1, wherein said file server change memory in-  
5 cludes a file server request.

6  
7           11.    A system as in claim 1, wherein said file server change memory is  
8 disposed to delay output to said mass storage element without delaying a response to file  
9 server requests.

10  
11           12.    A system as in claim 1, wherein  
12           said mass storage element includes a file storage system;  
13           each said file server is disposed for leaving said file storage system in an  
14 internally consistent state after processing file system changes;  
15           said internally consistent state is associated with a set of completed file  
16 system changes;  
17           said set of completed file system changes is identifiable by each said file  
18 server.

19  
20           13.    A system as in claim 1, wherein said mass storage element includes  
21 a file storage system and each said file server is disposed for leaving said file storage  
22 system in an internally consistent state after processing each said file system change.

1  
2 14. A file server system as in claim 1, wherein

3 said mass storage element includes a primary mass storage element and a

4 mirror mass storage element; and

5 said first file server processes said file system change for both said primary

6 mass storage element and said mirror mass storage element.

7  
8 15. A system as in claim 1, wherein said means for copying includes ac-

9 cess to at least one of said first and second file server change memories using a NUMA

10 network.

11  
12 16. A system as in claim 1, wherein said means for copying includes

13 remote memory access to at least one of said first and second file server change memo-

14 ries.

15  
16 17. A system as in claim 1, wherein said means for said second file

17 server to perform a file server request in its file server change memory is also operative in

18 response to a service interruption by said second file server.

19  
20 18. A file server system including

21 a first file server coupled to a first set of mass storage devices;

22 a second file server coupled to a second set of mass storage devices;

1 a server change memory;

2 said first file server disposed for receiving a file server request and in re-  
3 sponse thereto copying a descriptor of a file system change into said server change mem-  
4 ory; and

5 said first file server disposed for processing said file system change for both  
6 said first set of mass storage devices and for at least one said mass storage device in said  
7 second set.

8  
9 19. A system as in claim 18, wherein

10 said second file server is disposed for receiving a file server request and in  
11 response thereto copying a descriptor of a file system change into said server change  
12 memory; and

13 said second file server is disposed for processing said file system change  
14 for both said second set of mass storage devices and for at least one said mass storage de-  
15 vice in said first set.

16  
17 20. A system as in claim 18, wherein said server change memory in-  
18 cludes a disk block.

19  
20 21. A system as in claim 18, wherein said server change memory in-  
21 cludes a file server request.

1           22.    A system as in claim 18, wherein said server change memory in-  
2   cludes a first portion disposed at said first file server and a second portion disposed at  
3   said second file server.

4  
5           23.    A system as in claim 18, wherein

6           said server change memory includes a first portion disposed at said first file  
7   server and a second portion disposed at said second file server; and

8           said first file server is disposed for copying said descriptor into both said  
9   first portion and said second portion.

10  
11           24.    A system as in claim 18, wherein

12           said server change memory includes a first portion disposed at said first file  
13   server and a second portion disposed at said second file server; and

14           said first file server and said second file server are each disposed for copy-  
15   ing said descriptor into both said first portion and said second portion.

16  
17           25.    A system as in claim 18, wherein said server change memory is dis-  
18   posed to delay output to said mass storage element without delaying a response to file  
19   server requests.

20  
21           26.    A file server system including

1 a plurality of file servers, said plurality of file servers coupled to a mass  
2 storage element and at least one file server change memory;

3 each said file server disposed for receiving a file server request and in re-  
4 sponse thereto copying a descriptor of a file system change into said file server change  
5 memory; and

6 each said file server disposed for responding to a service interruption by  
7 performing a file system change in said file server change memory.

8

9 27. A system as in claim 26, including at least one said mass storage  
10 element for each said file server.

11

12 28. A system as in claim 26, including at least one said server change  
13 memory for each said file server.

14

15 29. A system as in claim 26, wherein a first said file server is disposed  
16 for processing said file system changes atomically, whereby a second said file server can  
17 on failover process exactly those file system changes not already processed by said first  
18 file server.

19

20 30. A system as in claim 26, wherein a first said file server is disposed to  
21 respond identically to service interruptions for itself and for a second said file server.

22

1           31.    A system as in claim 26, wherein at least one said file server delays  
2   output to said mass storage element without delaying a response to file server requests.

3  
4           32.    A system as in claim 26, wherein at least one said file server re-  
5   sponds to a file system change before committing a result of said file system change to  
6   mass storage.

7  
8           33.    A system as in claim 26, wherein  
9           each one of said file servers is coupled to at least a portion of said file  
10   server change memory using local memory access; and  
11           each one of said file servers is coupled to at least a portion of said file  
12   server change memory using remote memory access.

13  
14           34.    A system as in claim 26, wherein each said file server is disposed for  
15   copying said descriptors using a NUMA network.

16  
17           35.    A system as in claim 26, wherein each said file server is disposed for  
18   copying said descriptors using remote memory access.

19  
20           36.    A system as in claim 26, wherein said file server change memory  
21   includes a disk block.

22



1           37.    A system as in claim 26, wherein said file server change memory  
2 includes a file server request.

3  
4           38.    A system as in claim 26, wherein said file server change memory is  
5 disposed to delay output to said mass storage element without delaying a response to file  
6 server requests.

7  
8           39.    A system as in claim 26, wherein said mass storage element includes  
9 a file storage system and each said file server is disposed for leaving said file storage  
10 system in an internally consistent state after processing each said file system change.

11  
12           40.    A system as in claim 26, wherein  
13 said mass storage element includes a file storage system;  
14 each said file server is disposed for leaving said file storage system in an  
15 internally consistent state after processing file system changes;  
16 said internally consistent state is associated with a set of completed file  
17 system changes;

18           said set of completed file system changes is identifiable by each said file  
19 server.

20  
21           41.    A file server system as in claim 26, wherein

1           said mass storage element includes a primary mass storage element and a  
2 mirror mass storage element; and

3           said first file server processes said file system change for both said primary  
4 mass storage element and said mirror mass storage element.

5  
6           42.   A method of operating a file server system, said method including  
7 steps for

8           responding to an incoming file server request by copying a descriptor of a  
9 file system change to both a first file server and a second file server;

10          processing said file system change at said first file server while maintaining  
11 said descriptor copy at said second file server; and

12          performing, at said second file server, a file system change in response to a  
13 copied descriptor and a service interruption by said first file server.

14  
15          43.   A method as in claim 42, including steps for associating a first file  
16 server and a second file server with a mass storage element.

17  
18          44.   A method as in claim 42, including steps for delaying output by at  
19 least one said file server to said mass storage system without delaying a response to file  
20 system changes.

1           45.    A method as in claim 42, wherein a first said file server is disposed  
2 for processing said file system changes atomically, whereby a second said file server can  
3 on failover process exactly those file system changes not already processed by said first  
4 file server.

5  
6           46.    A method as in claim 42, wherein a first said file server is disposed  
7 to respond identically to service interruptions for itself and for a second said file server.

8  
9           47.    A method as in claim 42, wherein at least one said file server re-  
10 sponds to a file system change before committing a result of said file system change to  
11 mass storage.

12  
13           48.    A method as in claim 42, wherein  
14 each said file server includes a file server change memory;  
15 each one of said file servers is coupled to at least a portion of said file  
16 server change memory using local memory access; and  
17 each one of said file servers is coupled to at least a portion of said file  
18 server request memory using remote memory access.

19  
20           49.    A method as in claim 42, wherein said file server change memory  
21 includes a disk block.

1           50.    A method as in claim 42, wherein said file server change memory  
2 includes a file server request.

3  
4           52.    A method as in claim 42, wherein said file server change memory is  
5 disposed to delay output to said mass storage element without delaying a response to file  
6 server requests.

7  
8           53.    A method as in claim 42, wherein said mass storage element in-  
9 cludes a file storage system and each said file server is disposed for leaving said file stor-  
10 age system in an internally consistent state after processing each said file system change.

11  
12           54.    A method as in claim 42, wherein said steps for performing a file  
13 system change in response to a copied descriptor are also operative in response to a serv-  
14 ice interruption by said second file server.

15  
16           55.    A method as in claim 42, wherein said steps for processing includes  
17 steps for processing said file system change at both a primary mass storage element and a  
18 mirror mass storage element.

19  
20           56.    A method of operating a file server system, said method including  
21 steps for

1 receiving a file server request at one of a plurality of file servers and in re-  
2 sponse thereto copying a descriptor of a file system change into a server change memory;

3 processing said file system change for both a first set of mass storage de-  
4 vices coupled to a first one said file server and for at least one said mass storage device in  
5 a second set of mass storage devices coupled to a second one said file server.

6  
7 <sup>56</sup> 57. A method as in claim 56, wherein said descriptor includes a file  
8 server request.

9  
10 <sup>57</sup> 58. A method as in claim 56, wherein said server change memory in-  
11 cludes a disk block.

12  
13 <sup>58</sup> 59. A method as in claim 56, wherein said server change memory in-  
14 cludes a file server request.

15  
16 <sup>59</sup> 60. A method as in claim 56, wherein said server change memory in-  
17 cludes a first portion disposed at said first file server and a second portion disposed at  
18 said second file server.

19  
20 <sup>60</sup> 61. A method as in claim 56, wherein said server change memory in-  
21 cludes a first portion disposed at said first file server and a second portion disposed at

1 said second file server; and wherein said steps for copying include steps for copying said  
 2 descriptor into both said first portion and said second portion.

3

4 <sup>62</sup> 62. A method as in claim 56, wherein said server change memory in-  
 5 cludes a first portion disposed at said first file server and a second portion disposed at  
 6 said second file server; and said steps for copying include steps for copying said descrip-  
 7 tor into both said first portion and said second portion by either of said first file server or  
 8 said second file server.

9

10 <sup>63</sup> 63. A method as in claim 56, wherein said server change memory is dis-  
 11 posed to delay output to said mass storage element without delaying a response to file  
 12 server requests.

13

14 <sup>64</sup> 64. A method as in claim 56, wherein  
 15 said steps for receiving include receiving a file server request at either said  
 16 first file server or said second file server, and said steps for copying said descriptor in-  
 17 clude copying by either said first file server or said second file server; and including steps  
 18 for

19 processing said file system change for both said second set of mass storage  
 20 devices and for at least one said mass storage device in said first set.

21

64  
~~65~~

1 A method of operating a file server system, said method including  
2 steps for  
3 receiving a file server request at one of a plurality of file servers and in re-  
4 sponse thereto copying a descriptor of a file system change into a file server change  
5 memory; and  
6 responding to a service interruption by performing a file system change in  
7 response to a descriptor in said file server change memory.

65  
~~66~~

9 A method as in claim 65, including steps for associating a plurality  
10 of file servers with at least one mass storage element and at least one file server change  
11 memory.

66  
~~67~~

13 A method as in claim 65, including steps for delaying output to a  
14 mass storage element without delaying a response to file server requests.

67  
~~68~~

16 A method as in claim 65, including steps for leaving a file storage  
17 system on said mass storage element in an internally consistent state after processing  
18 each said file system change.

68  
~~69~~

20 A method as in claim 65, including steps for  
21 leaving a file storage system on said mass storage element in an internally  
22 consistent state after processing file system changes;

1 associating said internally consistent state with a set of completed file sys-  
 2 tem changes; and

3 identifying said set of completed file system changes by at least one said  
 4 file server.

5 <sup>69</sup>  
 6 ~~70.~~ A method as in claim 65, including steps for performing said re-  
 7 ceived file server request at both a primary mass storage element and a mirror mass stor-  
 8 age element.

9  
 10 <sup>70</sup>~~71.~~ A method as in claim 65, including steps for  
 11 processing said file system changes atomically at a first said file server; and  
 12 on failover processing exactly those file system changes not already proc-  
 13 essed by said first file server.

14  
 15 <sup>71</sup>~~72.~~ A method as in claim 65, including steps for responding identically  
 16 at a first said file server to service interruptions for itself and for a second said file server.

17  
 18 <sup>72</sup>~~73.~~ A method as in claim 65, wherein said file server change memory  
 19 includes a disk block.

20  
 21 <sup>73</sup>~~74.~~ A method as in claim 65, wherein said file server change memory  
 22 includes a file server request.



1

2

74  
75.

3

4

A method as in claim 65, wherein said file server change memory is disposed to delay output to said mass storage element without delaying a response to file server requests.

5

6

15  
76.

7

8

A method as in claim 65, including steps for responding to a file system change before committing a result of said file system change to mass storage at one said file server.